Approximate Fisher Kernels of non-iid Image Models for Image Categorization

Ramazan Gokberk Cinbis, Jakob Verbeek, Cordelia Schmid

Approximate Fisher Kernels of non-iid Image Models for Image Categorization

Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, *Fellow, IEEE*

Abstract—The bag-of-words (BoW) model treats images as sets of local descriptors and represents them by visual word histograms. The Fisher vector (FV) representation extends BoW, by considering the first and second order statistics of local descriptors. In both representations local descriptors are assumed to be identically and independently distributed (iid), which is a poor assumption from a modeling perspective. It has been experimentally observed that the performance of BoW and FV representations can be improved by employing discounting transformations such as power normalization. In this paper, we introduce non-iid models by treating the model parameters as latent variables which are integrated out, rendering all local regions dependent. Using the Fisher kernel principle we encode an image by the gradient of the data log-likelihood w.r.t. the model hyper-parameters. Our models naturally generate discounting effects in the representations; suggesting that such transformations have proven successful because they closely correspond to the representations obtained for non-iid models. To enable tractable computation, we rely on variational free-energy bounds to learn the hyper-parameters and to compute approximate Fisher kernels. Our experimental evaluation results validate that our models lead to performance improvements comparable to using power normalization, as employed in state-of-the-art feature aggregation methods.

Index Terms—Statistical image representations, object recognition, image classification, Fisher kernels.

1 Introduction

PATCH-based image representations, such bag of visual words (BoW) [10], [49], are widely utilized in image categorization and retrieval systems. BoW descriptor represents an image as a histogram over visual word counts. The histograms are constructed by mapping local feature vectors in images to cluster indices, where the clustering is typically learned using k-means. Perronnin and Dance [38] have enhanced this basic representation using the notion of Fisher kernels [20]. In this case local descriptors are soft-assigned to components of a mixture of Gaussian (MoG) density, and the image is represented using the gradient of the log-likelihood of the local descriptors w.r.t. the MoG parameters. As we show below, both BoW as well as MoG Fisher vector representations are based on models that assume that local descriptors are independently and identically distributed (iid). However, the iid assumption is a very poor one from a modeling perspective, see the illustration in Figure 1.

In this work, we consider models that capture the dependencies among local image regions by means of non-iid but completely exchangeable models, *i.e.* like iid models our models still treat the image as an unordered set of regions. We treat the parameters of the BoW models as latent variables with prior distributions learned from data. By integrating out the latent variables, all image regions become mutually dependent. We generate image representations from these models by applying the Fisher kernel principle, in this case by taking the gradient of the log-likelihood of the data in an image w.r.t. the hyper-parameters that control the priors on the latent model parameters.

Fig. 1. Local image appearance is not iid: the visible regions are informative on the masked-out ones; one has the impression to have seen the complete image by looking at half of the pixels.

However, in some cases, the gradient of the log-likelihood of the data can be intractable to compute. To compute a gradient-based representation in such cases, we replace the intractable log-likelihood with a tractable variational bound. We then compute gradients with respect to this bound instead of the likelihood. Following [4], which is the first and one of the very few studies utilizing this approximation method, we refer to the resulting kernel as the *variational Fisher kernel*. We show that the variational Fisher kernel is equivalent to the actual Fisher kernel when the variational bound is tight. Therefore, the variational Fisher kernel provides not only a technique for approximating intractable Fisher kernels, but also an alternative formulation for computing exact Fisher kernels. We demonstrate through examples that the variational formulation can be mathematically more convenient for deriving Fisher vectors representations.

- *R. G. Cinbis is with Milsoft, Ankara, Turkey. Most of the work in this paper was done when he was with the LEAR team, Inria Grenoble, France. E-mail: firstname.lastname@inria.fr*
- *J. Verbeek and C. Schmid are with LEAR team, Inria Grenoble Rhône-Alpes, Laboratoire Jean Kuntzmann, CNRS, Univ. Grenoble Alpes, France. E-mail: firstname.lastname@inria.fr*

In this work, we analyze three non-iid image models. Our first model is the multivariate Pólya model which represents the set of visual word indices of an image as independent draws from an unobserved multinomial distribution, itself drawn from a Dirichlet prior distribution. By integrating out the latent multinomial distribution, a model is obtained in which all visual word indices are mutually dependent. Interestingly, we find that our non-iid models yield gradients that are qualitatively similar to popular transformations of BoW image representations, such as square-rooting histogram entries or more generally applying power normalization [22], [39], [40], [52]. Therefore, our first contribution is to show that such transformations appear naturally if we remove the unrealistic iid assumption, *i.e.*, to provide an explanation why such transformations are beneficial.

Our second contribution is the analysis of Fisher vector representations over the latent Dirichlet allocation (LDA) model [3] for image classification purposes. The LDA model can capture the co-occurrence statistics missing in BoW representations. In this case the computation of the gradients is intractable, therefore, we compute approximate variational Fisher vectors [4]. We compare performance to Fisher vectors of PLSA [19], a topic model that does not treat the model parameters as latent variables. We find that topic models improve over BoW models, and that the LDA improves over PLSA even when square-rooting is applied.

Our third contribution is our most advanced model, which assumes that the local descriptors are iid samples from a latent MoG distribution, and we integrate out the mixing weights, means and variances of the MoG distribution. Since the computation of the gradients is intractable, we also use the variational Fisher kernel framework for this model. This leads to a representation that performs on par with the improved Fisher vector (FV) representation of [40] based on iid MoG models, which includes power normalization.

In our experimental analysis, we present a detailed experimental evaluation of the proposed the non-iid image models over local SIFT descriptors. In addition, we demonstrate that the latent MoG image model can effectively be combined with Convolutional Neural Network (CNN) based features. We consider two approaches for this purpose. First, following recent work [16], [31], we compute Fisher vectors over densely sampled image patches that are encoded using CNN features. Second, we propose to extract Fisher vectors over image regions sampled by a selective search method [50]. The experimental results on the PASCAL VOC 2007 [14] and MIT Indoor Scenes [41] datasets confirm the effectiveness of the proposed latent MoG image model, and the corresponding non-iid image descriptors.

This paper extends our earlier paper [8]. We present more complete and detailed discussions of related work and the variational Fisher kernel framework. We give a proof that Fisher kernels given by the traditional form and the variational framework are equivalent when the variational bound is tight. We extend the experimental evaluation of the proposed non-iid MoG models by evaluating them over CNN-based local descriptors. We also show that the classification results can be further improved by computing the CNN features over selective search windows, compared to using densely sampled image regions. We perform additional experimental evaluation on the MIT Indoor dataset [41]. Finally, we present a new empirical study on the relationship between the model likelihood and image categorization performance.

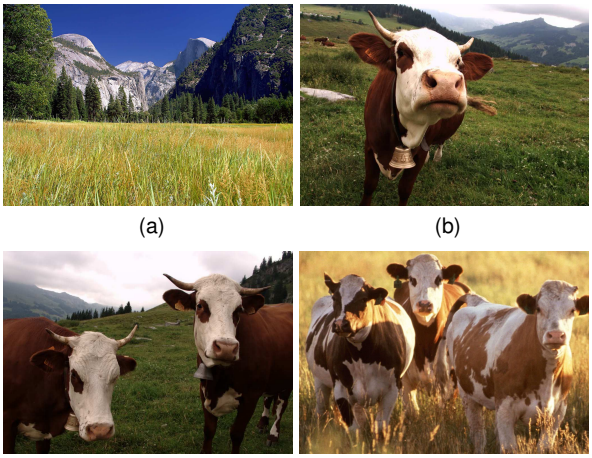


Fig. 2. The score of a linear 'cow' classifier will increase similarly from images (a) through (d) due to the increasing number of cow patches. This is undesirable: the score should sharply increase from (a) to (b), and remain stable among (b), (c), and (d).

2 Related Work

The use of non-linear feature transformations in BoW image representations is widely recognized to be beneficial for image categorization [22], [39], [40], [52], [56]. These transformations alleviate an obvious shortcoming of linear classifiers on BoW image representations: the fact that a fixed change Δ in a BoW histogram, from h to $h+\Delta$, leads to a score increment that is independent of the original histogram h: $f(h+\Delta) - f(h) = w^\top(h+\Delta) - w^\top h = w^\top \Delta$. This means that the effect on the score for a change Δ is not dependent on the context h in which it appears. Therefore, the score increment from images (a) though (d) in Figure 2 will be comparable, which is undesirable: the classifier score for *cow* should sharply increase from (a) to (b), and then remain stable among (b), (c), and (d).

Popular remedies to this problem include the use of chi-square kernels [56], or taking the square-root of histogram entries [39], also referred to as the Hellinger kernel [52]. Power normalization [39], defined as $f(x) = sign(x)|x|^\rho$, is a similar transformation that can be applied to non-histogram feature vectors, and it is equivalent to signed square-rooting for the coefficient $\rho = 1/2$. The effect of all of these is similar: they transform the features such that the first few occurrences of visual words will have a more pronounced effect on the classifier score than if the count is increased by the same amount but starting at a larger value. This is desirable, since now the first patches providing evidence for an object category can significantly impact the score, and hence making it for example easier to detect small object instances. The qualitative similarity is illustrated in Figure 3, where we compare the ℓ_2, chi-square, and Hellinger distances on the range $[0, 1]$.

The motivation for these transformations tends to vary in the literature. Sometimes it is based on empirical observations of improved performance [39], [52], by reducing sparsity in Fisher vectors [40], or in terms of variance stabilizing transformations [22], [55]. Recently, Kobayashi [26] showed that a similar discounting transformation based on taking logarithm of histogram entries, can be derived via modeling ℓ_1-normalized descriptors by Dirichlet distribution. Rana *et al*. [43] propose to discriminatively learn power normalization coefficients for image retrieval using a triplet-

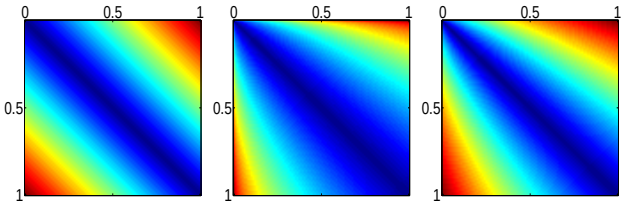


Fig. 3. Comparison of ℓ_2, Hellinger, and chi-square distances for values in the unit interval. Both the Hellinger and chi-square distance discount the effect of small changes in large values, unlike the ℓ_2 distance.

based objective function, which aims to obtain smaller distances across matching image pairs than non-matching ones. In contrast to these studies, we show that such discounting transformations appear naturally in generative image models that avoid making the unrealistic iid assumption that underlies the standard BoW and MoG-FV image representations.

Similar transformations are also used in image retrieval to counter burstiness effects [21], *i.e*., if rare visual words occur in an image, they tend to do so in bursts due to the locally repetitive nature of natural images. Burstiness also occurs in text, and the Dirichlet compound multinomial distribution, also known as multivariate Pólya distribution, has been used to model this effect [33]. This model places a Dirichlet prior on a latent per-document multinomial, and words in a document are sampled independently from it. Elkan [13] shows the relationship between the Fisher kernel of the multivariate Pólya distribution and the tf-idf document representation. In Section 4, we investigate the Fisher kernel based on multivariate Pólya distribution as our most basic non-iid image representation.

Our use of latent Dirichlet allocation (LDA) [3] differs from earlier work on using topic models such as LDA or PLSA [19] for object recognition [29], [42]. The latter use topic models to compress BoW image representations by using the inferred document-specific topic distribution. Similarly, Chandalia and Beal [4] propose to compress BoW document representation by computing LDA Fisher vector with respect to the parameters of the Dirichlet prior on the topic distributions. We, instead, use the Fisher kernel framework to expand the image representation by decomposing the original BoW histogram into several bags-of-words, one per topic, so that individual histogram entries not only encode how often a word appears, but also in combination with which other words it appears. Whereas compressed topic model representations were mostly found to at best maintain BoW performance, we find significant gains by using topic models. Finally, in contrast to the PLSA Fisher kernel, which was previously studied as a document similarity measure [5], [18], we show that the LDA Fisher kernel naturally involves discounting transformations.

Several other generative models have been proposed to capture spatial regularities across image regions. For example, the Spatial LDA model [53] extends the LDA model such that spatially neighboring visual words are more likely assigned to the same topic. The counting grid model [36], which is a grid of multinomial distributions, can be considered as an alternative to the spatial topic models. In this approach, the visual words of an image are treated as samples from a latent local neighborhood of the counting grid. Therefore, each local neighborhood of the model can be interpreted as a spatial grid of topics. While these studies show that incorporation of spatial information can improve unsupervised semantic segmentation results [53], or lead to better generative classifiers compared to LDA [36], we limit our focus to Fisher kernels of orderless, *i.e*. exchangeable, generative models in our study.

The computation of the LDA Fisher vector image representation is technically more involved compared to the PLSA model. In the case of the LDA model, the latent model parameters cannot be integrated out analytically, and the computation of the gradients is no longer tractable. Similarly, the Fisher kernel for our Latent MoG image model is intractable since the latent variables (mixing weights, means, and variances) cannot be integrated out analytically. We overcome this difficulty by relying on the variational free-energy bound [24], which is obtained by subtracting the Kullback-Leibler divergence between an approximate posterior on the latent variables and the true posterior. By imposing a certain independence structure on the approximate posterior, tractable approximate inference techniques can be devised. We then compute the gradient of the variational bound as a surrogate for the intractable gradients of the exact log-likelihood. The method of approximating Fisher kernels with the gradient vector of a variational bound was first proposed by Chandalia and Beal [4] in order to obtain the LDA Fisher kernel. The only other work incorporating this technique, to the best of our knowledge, is the recent work of Perina *et al*. [37], which proposes a variational Fisher kernel for micro-array data. We show that variational Fisher kernel is equivalent to the exact Fisher vector when the variational bound is tight, and demonstrate that in some cases it can be a mathematically more convenient formulation, compared to the original Fisher kernel definition. Finally, we note that the variational approximation method for Fisher kernels differs from Perina *et al*. [35], which uses the variational free-energy to define an alternative encoding, replacing the Fisher kernel.

In the following section we review the Fisher kernel framework, and the variational approximation method. In Section 4 we present our non-iid latent variable models and propose novel Fisher vector representations based on them. We present experimental results in Section 5, and summarize our conclusions in Section 6.

3 Fisher vectors and variational approximation

In this section we present an overview of the Fisher kernel framework, variational inference, and the variational Fisher kernel.

3.1 Fisher vectors

Images can be considered as samples from a generative process, and therefore class-conditional generative models can be used for image categorization. However, it is widely observed that discriminative classifiers typically outperform classification based on generative models, see *e.g*. [17]. A simple explanation is that discriminative classifiers aim to maximize the end goal, which is to categorize entities based on their content. In contrast, generative classifiers instead require modeling class-conditional data distributions, which is arguably a more difficult task than only learning decision surfaces, and therefore result in inferior categorization performance.

The Fisher kernel framework proposed by Jaakkola and Haussler [20] allows combining the power of generative models and

discriminative classifiers. In particular, Fisher kernel provides a framework for deriving a kernel from a probabilistic model. Suppose that $p(\mathbf{x})$ is a generative model with parameters θ.[1] Then, the Fisher kernel $K(\mathbf{x}, \mathbf{x}')$ is defined as

$$K(\mathbf{x}, \mathbf{x}') = g(\mathbf{x})^{\mathrm{T}} I^{-1} g(\mathbf{x}') \,, \tag{1}$$

where the gradient $g(\mathbf{x}) = \nabla_\theta \log p(\mathbf{x})$ is called the *Fisher score*, and I is the Fisher information matrix:

$$I = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[g(\mathbf{x})g(\mathbf{x})^{\mathrm{T}}\right]. \tag{2}$$

which is equivalent to the covariance of the Fisher score as computed using $p(\mathbf{x})$, since $\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[g(\mathbf{x})] = \mathbf{0}$. The inner product space (*i.e.* explicit feature mapping) induced by a Fisher kernel is given by

$$\phi(\mathbf{x}) = I^{-\frac{1}{2}} g(\mathbf{x}), \tag{3}$$

where $I^{-\frac{1}{2}}$ is the whitening transform using the Fisher information matrix. Sánchez *et al*. [45] refer to the normalized gradient given by $\phi(\mathbf{x})$ as the *Fisher vector*. In practice the term "Fisher vector" is sometimes also used to refer to the non-normalized gradients (*i.e.* Fisher score) as well.

The essential idea in Fisher kernel is to use gradients $g(\mathbf{x})$ of the data log-likelihood to extract features w.r.t. a generative model. The Fisher information matrix, on the other hand, is of lesser importance. A theoretical motivation for using I is that $I^{-1}g(\mathbf{x})$ gives the steepest descent direction along the manifold of the parameter space, which is also known as the *natural gradient*. Another motivation is that I makes the Fisher kernel invariant to the re-parameterization $\theta \rightarrow \psi(\theta)$ for any differentiable and invertible function ψ [2], which can be easily shown using the chain rule and the Jakobian matrix of the inverse function ψ^{-1}.

However, the computation of the Fisher information matrix I is intractable for many models. Although in principle it can be approximated empirically as $I \approx \frac{1}{|X|}\sum_{\mathbf{x}\in X} g(\mathbf{x})g(\mathbf{x})^{\mathrm{T}}$, the approximation itself can be costly if $g(\mathbf{x})$ is high dimensional. In such cases, empirical approximation can be used only for the diagonal terms. Alternatively, I can be dropped altogether [20] or analytical approximations can be derived, see *e.g.* [38], [45], [46].

3.2 Variational approximate inference

Variational methods are a family of mathematical tools that can be used to approximate intractable computations, particularly those involving difficult integrals. Originally developed in statistical physics based on the *calculus of variations* and the *mean field theory*, the variational approximation framework that we utilize in this paper is known as the *variational inference*, and it is now among the most successful approximate probabilistic inference techniques [2], [24], [32].

In the context of probabilistic models, the central idea in variational methods is to devise a bound on the log-likelihood function in terms of an approximate posterior distribution over the latent variables. Let X denote the set of observed variables, and Λ denote the set of latent variables and latent parameters. Suppose that $q(\Lambda)$ is an approximate distribution over the latent variables. Then, the distribution $p(X)$ can be decomposed as follows for any choice of the approximate posterior q:

$$\ln p(X) = F(p, q) + D\big(q||p\big). \tag{4}$$

1. We drop the model parameters θ from function arguments for brevity.

In this equation, F is the *variational free-energy* given by

$$F(p, q) = \int q(\Lambda) \ln\left(\frac{p(X, \Lambda)}{q(\Lambda)}\right) d\Lambda \tag{5}$$

$$= \mathbb{E}_{q(\Lambda)}[\ln p(X, \Lambda)] + H(q), \tag{6}$$

where $H(q)$ is the entropy of the distribution q. The term $D\left(q||p\right)$ in Eq. (4) is the Kullback-Leibler (KL) divergence between the distributions $q(\Lambda)$ and $p(\Lambda|X)$:

$$D(q||p) = -\int q(\Lambda) \ln\left(\frac{p(\Lambda|X)}{q(\Lambda)}\right) d\Lambda. \tag{7}$$

Since the KL-divergence term $D\left(q||p\right)$ is strictly non-negative, the variational free energy $F(p, q)$ is a lower-bound on the true log-likelihood $\ln p(X)$, *i.e.* $F(p, q) \leq \ln p(X)$. When the KL-divergence term is zero, *i.e.* the distribution q is equivalent to the true posterior distribution, the bound F is tight.

In order to effectively utilize the decomposition in Eq. (6) for a given distribution p, we need to choose the distribution q such that it leads to a tractable and as tight as possible lower-bound $F(p, q)$. For this purpose, we constrain q to a family of distributions $\mathcal{Q}$ that leads to tractable computations, typically by imposing independence assumptions. For example suppose that $\Lambda = (\lambda_1, \ldots, \lambda_n)$, we may choose $\mathcal{Q}$ to be the set of distributions that factorize over the λ_i, *i.e.* with $q(\Lambda) = \prod_{i=1}^{n} q_i(\lambda_i)$. Given the family $\mathcal{Q}$, we maximize $F(p, q)$ by minimizing the KL divergence in Eq. (4) over all $q \in \mathcal{Q}$.

3.3 Variational Fisher kernel

In this paper, we utilize the variational free-energy bounds for two purposes. The first is to estimate the hyper-parameters of the LDA (Section 4.2) and the Latent MoG (Section 4.3) models from training data using an approximate maximum likelihood procedure. For this purpose, we iteratively update the variational lower-bound with respect to the variational distribution parameters, and the model hyper-parameters; an approach that is known as the *variational expectation-maximization* procedure [24].

Our second main use of the variational free-energy is to compute approximate Fisher vectors where the original Fisher vector is intractable to compute. In particular, we approximate the Fisher vector by the gradient of the variational lower-bound given by Eq. (6), *i.e.* $g(\mathbf{x}) \approx \nabla_\theta F(p, q)$, which we refer to as *variational Fisher vector*. Since, the entropy $H(q)$ is constant w.r.t. model parameters, the variational Fisher vector θ_q can equivalently be written as

$$\phi_q(X) = I^{-\frac{1}{2}} \nabla_\theta \mathbb{E}_q[\ln p(X, \Lambda)]\,. \tag{8}$$

where I is the (approximate) Fisher information matrix.

We have already discussed that the variational bound in Eq. (6) is tight when the distribution q matches the posterior on the hyper-parameters. We will now show that its gradient equals that of the data log-likelihood if the bound is tight. In order to prove this, we first write the partial derivative of the lower-bound with respect to some model (hyper-)parameter θ:

$$\frac{\partial F}{\partial \theta} = \frac{\partial \mathbb{E}_q[\ln p(X, \Lambda)]}{\partial \theta}. \tag{9}$$

By definition, we can interchange the differential operator and the expectation:

$$\frac{\partial F}{\partial \theta} = \mathbb{E}_q\left[\frac{\partial \ln p(X, \Lambda)}{\partial \theta}\right]. \tag{10}$$

Without loss of generality, we assume that all latent variables are continuous, in which case the expectation is equivalent to

$$\frac{\partial F}{\partial \theta} = \int q(\Lambda) \frac{\partial \ln p(X,\Lambda)}{\partial \theta} d\Lambda. \quad (11)$$

By following differentiation rules, we obtain the equation:

$$\frac{\partial F}{\partial \theta} = \int q(\Lambda) \frac{1}{p(\Lambda|X)p(X)} \frac{\partial p(X,\Lambda)}{\partial \theta} d\Lambda. \quad (12)$$

Since the bound is assumed to be tight, the $q(\Lambda)$ and $p(\Lambda|X)$ are identical. In addition, we observe that $p(X)$ is a constant with respect to the integration variables. Therefore, we can simplify the equation as follows:

$$\frac{\partial F}{\partial \theta} = \frac{1}{p(X)} \int \frac{\partial p(X,\Lambda)}{\partial \theta} d\Lambda, \quad (13)$$

which can be re-written as follows:

$$\frac{\partial F}{\partial \theta} = \frac{1}{p(X)} \frac{\partial \int p(X,\Lambda) d\Lambda}{\partial \theta}. \quad (14)$$

Finally, we integrate out Λ and simplify the equation into the following form:

$$\frac{\partial F}{\partial \theta} = \frac{\partial \ln p(X)}{\partial \theta}, \quad (15)$$

which completes the proof.

In addition to presenting a relationship between the original Fisher vector and the variational Fisher vector definitions, the proof shows that the latter formulation can be used as an alternative framework. In fact, we observe that the variational formulation can in some cases be mathematically more convenient to derive Fisher vector representations. Even though our main interest in this paper is to compute approximate representations based on the LDA and latent MoG image models presented in the next section, we present two additional examples in Appendix A that demonstrate the usefulness of the variational formulation.

4 Non-iid Image Representations

In this section we present our non-iid models for local image descriptors. We start with a model for BoW quantization indices, and extend the model to capture co-occurrence statistics across visual words using LDA in Section 4.2. Finally, we consider a non-iid extension of mixture of Gaussian models over sets of local descriptors in Section 4.3.

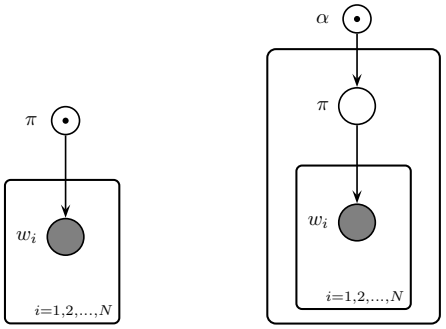
4.1 Bag-of-words and the multivariate Pólya model

The standard BoW image representation can be interpreted as applying the Fisher kernel framework to a simple iid multinomial model over visual word indices, as shown in [27]. Let $w_{1:N} = \{w_1, \ldots, w_N\}$ denote the visual word indices corresponding to N patches sampled in an image, and let π be a learned multinomial over K visual words, parameterized in log-space, *i.e.* $p(w_i = k) = \pi_k$ with $\pi_k = \exp(\gamma_k) / \sum_{k'} \exp(\gamma_{k'})$. The data likelihood for the BoW model is given by

$$p(w_{1:N}) = \prod_{i=1}^{N} p(w_i = k). \quad (16)$$

The gradient of the data log-likelihood is in this case given by

$$\frac{\partial \sum_{i=1}^{N} \ln p(w_i)}{\partial \gamma_k} = n_k - N\pi_k, \quad (17)$$



(a) Multinomial BoW model (b) Pólya model

Fig. 4. Graphical representation of the models in Section 4.1: (a) multinomial BoW model, (b) Pólya model. The outer plate in (b) refer to images. The index i runs over the visual word indices in an image. Nodes of observed variables are shaded, and those of (hyper-)parameters are marked with a central dot in the node.

where n_k denotes the number of occurrences of visual word k among the set of indices $w_{1:N}$. This is a shifted version of the standard BoW histogram, where the mean of all image representations is centered at the origin. We stress that this multinomial interpretation of the BoW model assumes that the visual word indices across all images are iid, which directly generates the product form in the likelihood of Eq. (16), and the count statistic in the gradient of the log-likelihood in Eq. (17).

Our first non-iid model assumes that for each image there is a different, a-priori unknown, multinomial generating the visual word indices in that image. In this model visual word indices within an image are mutually dependent, since knowing some of the w_i provides information on the underlying multinomial π, and thus also provides information on which subsequent indices could be sampled from it. The model is parameterized by a non-symmetric Dirichlet prior over the latent image-specific multinomial, $p(\pi) = \mathcal{D}(\pi|\alpha)$ with $\alpha = (\alpha_1, \ldots, \alpha_K)$, and the w_i are modeled as iid samples from π. The marginal distribution on the w_i is obtained by integrating out π:

$$p(w_{1:N}) = \int p(\pi) \prod_{i=1}^{N} p(w_i|\pi) d\pi. \quad (18)$$

This model is known as the multivariate Pólya, or Dirichlet compound multinomial [33], and the integral simplifies to

$$p(w_{1:N}) = \frac{\Gamma(\hat{\alpha})}{\Gamma(N+\hat{\alpha})} \prod_{k=1}^{K} \frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)}, \quad (19)$$

where $\Gamma(\cdot)$ is the Gamma function, and $\hat{\alpha} = \sum_{k=1}^{K} \alpha_k$. See Figure 4a and Figure 4b for a graphical representation of the BoW multinomial model, and the Pólya model.

Following the Fisher kernel framework, we represent an image by the gradient w.r.t. the hyper-parameter α of the log-likelihood of the visual word indices $w_{1:N}$. The partial derivative w.r.t. α_k is given by

$$\frac{\partial \ln p(w_{1:N})}{\partial \alpha_k} = \psi(\alpha_k + n_k) - \psi(\hat{\alpha} + N) - \psi(\alpha_k) + \psi(\hat{\alpha}), \quad (20)$$

where $\psi(x) = \partial \ln \Gamma(x) / \partial x$ is the digamma function.

Only the first two terms in Eq. (20) depend on the counts n_k, and for fixed N the gradient is determined up to additive constants

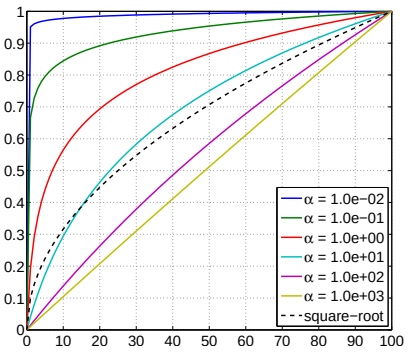


Fig. 5. Digamma functions $\psi(\alpha+n)$ for various α, and $\sqrt{n}$ as a function of n; functions have been rescaled to the range $[0,1]$.

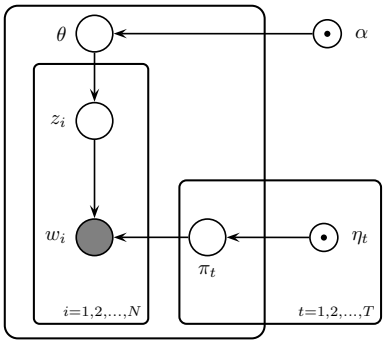


Fig. 6. Graphical representation of LDA. The outer plate refers to images. The index i runs over patches, and index t over topics.

by $\psi(\alpha_k + n_k)$, *i.e.* it is given by a transformation of the visual word counts n_k. Figure 5 shows the transformation $\psi(\alpha+n)$ for various values of α, along with the square-root function used in the Hellinger distance for reference. We see that the same monotone-concave discounting effect is obtained as by taking the square-root of histogram entries. This transformation arises naturally in our latent variable model, and suggests that such transformations are successful *because* they correspond to a more realistic non-iid model, *c.f.* Figure 1.

Observe that in the limit of $\alpha \to \infty$ the transfer function becomes linear, since for large α the Dirichlet prior tends to a delta peak on the multinomial simplex and thus removes the uncertainty on the underlying multinomial, with an observed multinomial BoW model as its limit. In the limit of $\alpha \to 0$, corresponding to priors that concentrate their mass at sparse multinomials, the transfer function becomes a step function. This is intuitive, since in the limit of ultimately sparse distributions only one word will be observed, and its count no longer matters, we only need to know which word is observed to determine which α_k should be increased to improve the log-likelihood.

4.2 Capturing co-occurrence with topic models

The Pólya model is non-iid but it does not model co-occurrence across visual words, this can be seen from the posterior distribution $p(w=k|w_{1:N}) = \int p(w=k|\pi)p(\pi|w_{1:N})d\pi \propto n_k + \alpha_k$. The model just predicts to see more visual words of the type it has already seen before. In our second model, we extend the Pólya model to capture co-occurrence statistics of visual words using latent Dirichlet allocation (LDA) [3]. We model the visual words in an image as a mixture of T topics, encoded by a multinomial θ mixing the topics, where each topic itself is represented by a multinomial distribution π_t over the K visual words. We associate a variable z_i, drawn from θ, with each patch that indicates which topic was used to draw its visual word index w_i. We place Dirichlet priors on the topic mixing, $p(\theta) = \mathcal{D}(\theta|\alpha)$, and the topic distributions $p(\pi_t) = \mathcal{D}(\pi_t|\eta_t)$, and integrate these out to obtain the marginal distribution over visual word indices as:

$$p(w_{1:N}) = \iint p(\theta)p(\pi)\prod_{i=1}^{N} p(w_i|\theta,\pi)d\theta d\pi, \quad (21)$$

$$p(w_i=k|\theta,\pi) = \sum_{t=1}^{T} p(z_i=t|\theta)p(w_i=k|\pi_t). \quad (22)$$

See Figure 6 for a graphical representation of the model. Note that this model is equivalent to the Pólya model discussed above when there is only a single topic, *i.e.* for $T=1$.

Both the log-likelihood and its gradient are intractable to compute for the LDA model. As discussed in Section 3.3, however, we can resort to variational methods to compute a free-energy bound F using an approximate posterior. Here we use a completely factorized approximate posterior as in [3] of the form

$$q(\theta,\pi_{1:T},z_{1:N}) = q(\theta)\prod_{t=1}^{T} q(\pi_t)\prod_{i=1}^{N} q(z_i). \quad (23)$$

The update equations of the variational distributions $q(\theta) = \mathcal{D}(\theta|\alpha^*)$ and $q(\pi_t) = \mathcal{D}(\pi_t|\eta_t^*)$ to maximize the free-energy bound F are given by:

$$\alpha_t^* = \alpha_t + \sum_{i=1}^{N} q_{it}, \qquad \eta_{tk}^* = \eta_{tk} + \sum_{i:w_i=k} q_{it}, \quad (24)$$

where $q_{it} = q(z_i=t)$, which is itself updated according to $q_{it} \propto \exp[\psi(\alpha_t^*) - \psi(\hat{\alpha}^*) + \psi(\eta_{tk}^*) - \psi(\hat{\eta}_t^*)]$. These update equations can be applied iteratively to monotonically improve the variational bound.

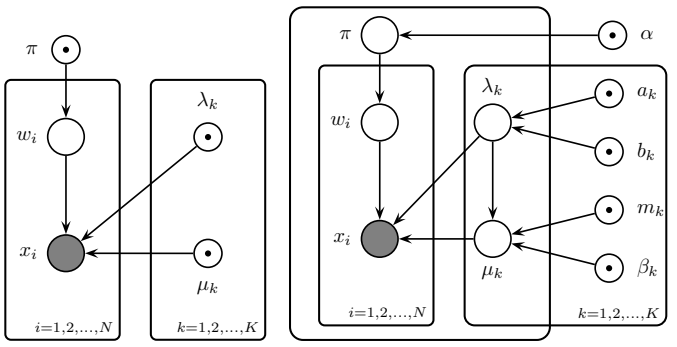
The gradients of F w.r.t. the hyper-parameters are obtained from these as

$$\frac{\partial F}{\partial \alpha_t} = \psi(\alpha_t^*) - \psi(\hat{\alpha}^*) - [\psi(\alpha_t) - \psi(\hat{\alpha})], \quad (25)$$

$$\frac{\partial F}{\partial \eta_{tk}} = \psi(\eta_{tk}^*) - \psi(\hat{\eta}_t^*) - [\psi(\eta_{tk}) - \psi(\hat{\eta}_t)]. \quad (26)$$

The gradient w.r.t. α encodes a discounted version of the topic proportions as they are inferred in the image. The gradients w.r.t. the hyper-parameters η_t can be interpreted as decomposing the bag-of-word histogram over the T topics, and encoding the soft counts of words assigned to each topic. The entries $\frac{\partial F}{\partial \eta_{tk}}$ in this representation not only code how often a word was observed but also in combination with which other words, since the co-occurrence of words throughout the image will determine the inferred topic mixing and thus the word-to-topic posteriors q_{it}.

In our experiments we compare LDA with the PLSA model [19]. This model treats the topics π_t, and the topic mixing θ as non-latent parameters which are estimated by maximum likelihood. To represent images using PLSA we apply the Fisher kernel framework and compute gradients of the log-likelihood w.r.t. θ and the π_t. The PLSA model with a single topic reduces to the iid multinomial model discussed in the previous section.



(a) MoG model (b) Latent MoG model

Fig. 7. Graphical representation of the models in Section 4.3: (a) MoG model, (b) latent MoG model. The outer plate in (b) without indexing refer to images. The index i runs over the local descriptors, and index k over Gaussians in the mixture which represent the visual words.

4.3 Modeling descriptors using latent MoG models

In this section we turn to the image representation of Perronnin and Dance [38] that applies the Fisher kernel framework to mixture of Gaussian (MoG) models over local descriptors. An improved version of this representation using power normalization was presented in [40].

A MoG density $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \sigma_k)$ is defined by mixing weights $\pi = \{\pi_k\}$, means $\mu = \{\mu_k\}$ and variances $\sigma = \{\sigma_k\}$.[2] The K Gaussian components of the mixture correspond to the K visual words in a BoW model. In [38], [40], local descriptors across images are assumed to be iid samples from a single MoG model underlying all images. They represent an image by the gradient of the log-likelihood of the extracted local descriptors $x_{1:N}$ w.r.t. the model parameters. Using the soft-assignments $p(k|x) = \pi_k \mathcal{N}(x; \mu_k, \sigma_k)/p(x)$ of local descriptors to mixture components the partial derivatives are computed as:

$$\frac{\partial \ln p(x_{1:N})}{\partial \gamma_k} = \sum_{i=1}^{N} p(k|x_i) - \pi_k, \tag{27}$$

$$\frac{\partial \ln p(x_{1:N})}{\partial \mu_k} = \sum_{i=1}^{N} p(k|x_i)(x - \mu_k)/\sigma_k, \tag{28}$$

$$\frac{\partial \ln p(x_{1:N})}{\partial \lambda_k} = \sum_{i=1}^{N} p(k|x_i) \left(\sigma_k - (x_i - \mu_k)^2\right)/2, \tag{29}$$

where we re-parameterize the mixing weights as $\pi_k = \exp(\gamma_k)/\sum_{k'=1}^{K} \exp(\gamma_{k'})$, and the Gaussians with precisions $\lambda_k = \sigma_k^{-1}$, as in [27]. For local descriptors of dimension D, the gradient yields an image representation of size $K(1+2D)$, since for each of the K visual words there is one derivative w.r.t. its mixing weight, and $2D$ derivatives for the means and variances in the D dimensions. This representation thus stores more information about the local descriptors assigned to a visual word than just their count, as a result higher recognition performance can be obtained using the same number of visual words as compared to the BoW representation.

In analogy to the Pólya model, we remove the iid assumption by defining a MoG model per image and treating its parameters as latent variables. We place conjugate priors on the image-specific parameters: a Dirichlet prior on the mixing weights, $p(\pi) = \mathcal{D}(\pi|\alpha)$, and a combined Normal-Gamma prior on the means μ_k and precisions $\lambda_k = \sigma_k^{-1}$:

$$p(\lambda_k) = \mathcal{G}(\lambda_k|a_k, b_k), \tag{30}$$

$$p(\mu_k|\lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k\lambda_k)^{-1}). \tag{31}$$

The distribution on the descriptors $x_{1:N}$ in an image is obtained by integrating out the latent MoG parameters:

$$p(x_{1:N}) = \iiint p(\pi)p(\mu,\lambda) \prod_{i=1}^{N} p(x_i|\pi,\mu,\lambda) d\pi d\mu d\lambda, \tag{32}$$

$$p(x_i|\pi,\mu,\lambda) = \sum_{k=1}^{K} p(w_i = k|\pi) p(x_i|w_i = k,\lambda,\mu), \tag{33}$$

where $p(w_i = k|\pi) = \pi_k$, and $p(x_i|w_i = k,\lambda,\mu) = \mathcal{N}(x_i|\mu_k, \lambda_k^{-1})$ is the Gaussian corresponding to the k-th visual word. See Figure 7a and Figure 7b for graphical representations of the MoG model and the latent MoG model.

Computing the log-likelihood in this model is also intractable, as is computing the gradient of the log-likelihood which we need for both hyper-parameter learning and to extract the Fisher vector representation. To overcome these problems we replace the intractable log-likelihood with its variational lower bound.

By constraining the variational posterior q in the bound F given by Eq. (6) to factorize as $q(\pi,\mu,\lambda,w_{1:N}) = q(\pi,\mu,\lambda)q(w_{1:N})$ over the latent MoG parameters and the assignments of local descriptors to visual words, we obtain a bound for which we can tractably compute its value and gradient w.r.t. the hyper-parameters. Given this factorization it is easy to show that the optimal q will further factorize as

$$q(\pi,\mu,\lambda,w_{1:N}) = q(\pi) \prod_{k=1}^{K} q(\mu_k|\lambda_k)q(\lambda_k) \prod_{i=1}^{N} q(w_i), \tag{34}$$

and that the variational posteriors on the model parameters will have the form of Dirichlet and Normal-Gamma distributions $q(\pi) = \mathcal{D}(\pi|\alpha^*)$, $q(\lambda_k) = \mathcal{G}(\lambda_k|a_k^*, b_k^*)$, $q(\mu_k|\lambda_k) = \mathcal{N}(\mu_k|m_k^*, (\beta_k^*\lambda_k)^{-1})$. Given the hyper-parameters we can update the variational distributions to maximize the variational lower bound. In order to write the update equations, it is convenient to define the following sufficient statistics :

$$s_k^0 = \sum_{i=1}^{N} q_{ik}, \quad s_k^1 = \sum_{i=1}^{N} q_{ik}x_i, \quad s_k^2 = \sum_{i=1}^{N} q_{ik}x_i^2. \tag{35}$$

where $q_{ik} = q(w_i = k)$. Then, the parameters of the optimal variational distributions on the MoG parameters for a given image are found as:

$$\alpha_k^* = \alpha_k + s_k^0, \tag{36}$$

$$\beta_k^* = \beta_k + s_k^0, \tag{37}$$

$$m_k^* = (s_k^1 + \beta_k m_k)/\beta_k^*, \tag{38}$$

$$a_k^* = a_k + s_k^0/2, \tag{39}$$

$$b_k^* = b_k + \frac{1}{2}(\beta_k m_k^2 + s_k^2) - \frac{1}{2}\beta_k^*(m_k^*)^2. \tag{40}$$

The component assignments $q(w_i)$ that maximize the bound given the variational distributions on the MoG parameters are given by:

$$\ln q_{ik} = \mathbb{E}_{q(\pi)q(\lambda_k,\mu_k)}\left[\ln \pi_k + \ln \mathcal{N}(x_i|\mu_k, \lambda_k^{-1})\right] \tag{41}$$

$$= \psi(\alpha_k^*) - \psi(\hat{\alpha}^*) + \frac{1}{2}\left[\psi(a_k^*) - \ln b_k^*\right] \tag{42}$$

$$-\frac{1}{2}\left[\frac{a_k^*}{b_k^*}(x_i - m_k^*)^2 + (\beta_k^*)^{-1}\right]. \tag{43}$$

2. We present here the uni-variate case for clarity, extension to the multi-variate case with diagonal covariance matrices is straightforward.

Since the sufficient statistics given by Eq. (35) depend on the component assignments, the distributions on the MoG parameters and the component assignments can be updated iteratively to improve the bound.

Using the above variational update equations, we obtain the variational distribution, and therefore the lower-bound on the log-likelihood for each image. During training, we learn the model hyper-parameters by iteratively maximizing the sum of the lower-bounds for the training images w.r.t. the hyper-parameters, and w.r.t. the variational parameters. Once the latent MoG model is trained, we use the per-image lower-bounds to extract the approximate Fisher vector descriptors according to the gradient of F with respect to the model hyper-parameters.

The gradient of F w.r.t. the hyper-parameters depends only on the variational distributions on the MoG parameters of an image $q(\pi)$, $q(\lambda_k)$, and $q(\mu_k|\lambda_k)$, and not on the component assignments $q(w_i)$. For the precision hyper-parameters we find:

$$\frac{\partial F}{\partial a_k} = [\psi(a_k^*) - \ln b_k^*] - [\psi(a_k) - \ln b_k]\,, \tag{44}$$

$$\frac{\partial F}{\partial b_k} = \frac{a_k}{b_k} - \frac{a_k^*}{b_k^*}, \tag{45}$$

For the hyper-parameters of the means:

$$\frac{\partial F}{\partial \beta_k} = \frac{1}{2}\left(\beta_k^{-1} - \frac{a_k^*}{b_k^*}(m_k - m_k^*)^2 - 1/\beta_k^*\right), \tag{46}$$

$$\frac{\partial F}{\partial m_k} = \beta_k \frac{a_k^*}{b_k^*}(m_k^* - m_k), \tag{47}$$

and for the hyper-parameters of the mixing weights:

$$\frac{\partial F}{\partial \alpha_k} = [\psi(\alpha_k^*) - \psi(\hat{\alpha}^*)] - [\psi(\alpha_k) - \psi(\hat{\alpha})]\,. \tag{48}$$

By substituting the update equation (36) for the variational parameters α_k^* in the gradient Eq. (48), we exactly recover the gradient of the multivariate Pólya model, albeit using soft-counts $s_k^0 = \sum_{i=1}^N q(w_i = k)$ of visual word occurrences here. Thus, the bound leaves the qualitative behavior of the multivariate Pólya model intact. Similar discounting effects can be observed in the gradients of the hyper-parameters of the means and variances. Substitution of the update equation (38) for the variational parameters m_k^* in the gradient Eq. (47), reveals that the gradient is similar to the square-root of the gradient obtained in [38] for the MoG mean parameters. The discounting function for this gradient is however slightly different from the $\psi(\cdot)$ function, but has a similar monotone concave form. We consider examples of the learned discounting functions in Section 5.4.

Our latent MoG model associates two hyper-parameters (m_k, β_k) with each mean μ_k, and similar for the precisions. Therefore, our image representation are almost twice as long compared to the iid MoG model: $K(1+4D)$ *vs.* $K(1+2D)$ dimensions. The updates of the variational parameters β_k^* and a_k^* in equations (37) and (39), however, only involve the zero-order statistics s_k^0. In [38] the FV components corresponding to the mixing weights of the MoG, which are also based on zero-order statistics, were shown to be redundant when also including the components corresponding to the means and variances. Therefore, we expect the gradients w.r.t. the corresponding hyper-parameters β_k and a_k to be of little importance for image classification purposes. Experimental results, not reported here, have empirically verified this. We therefore fix the number of Gaussians rather than the FV dimension when we compare different representations in the next section, and use all FV components to avoid confusion.

5 Experimental evaluation

In this section, we present a detailed evaluation of the latent BoW, LDA and the latent MoG models over SIFT local descriptors using the PASCAL VOC'07 [14] data set in Section 5.2, Section 5.3 and Section 5.4, respectively. Then, we present a empirical study on the relationship between the model likelihood and image categorization performance in Section 5.5. Finally, we evaluate the Latent MoG model, which is the most advanced model that we consider, over the CNN-based local descriptors, and compare against the state-of-the-art on the PASCAL VOC'07 and MIT Indoor [41] data sets in Section 5.6.

Now, we first describe our experimental setup for the SIFT-based experiments used in the subsequent sections.

5.1 Experimental setup

In order to extract SIFT descriptors, we use the experimental setup described in the evaluation paper of Chatfield *et al*. [6]: we sample local SIFT descriptors from the same dense grid (3 pixel stride, across 4 scales), which results in around $60,000$ patches per image, project the local descriptors to 80 dimensions with PCA, and train the MoG visual vocabularies from 1.5×10^6 descriptors. For the PASCAL VOC'07 data set, we use the interpolated mAP score specified by the VOC evaluation protocol [14].

We compare global image representations, and representations that capture spatial layout by concatenating the signatures computed over various spatial cells as in the spatial pyramid matching (SPM) method [30]. Again, we follow [6] and combine a 1×1, a 2×2, and a 3×1 grid. Throughout, we use linear SVM classifiers, and we cross-validate the regularization parameter.

Before training the classifiers we apply two normalizations to the image representations. First, we whiten the representations so that each dimension is zero-mean and has unit-variance across images in order to approximate normalization with the inverse Fisher information matrix. Second, following [40], we also ℓ_2 normalize the image representations.

For the BoW, PLSA and MoG models, we compare using Fisher vectors with and without power normalization, and to using the Fisher vectors of the corresponding latent variable models. As in [40], power normalization is applied after whitening, and before ℓ_2 normalization. We evaluate two types of power normalization: (i) signed square-rooting ($\rho = 1/2$) as in [6], [40], which we denote by a prefix "Sqrt", (ii) more general power normalization, which we denote by a prefix "Pn". In the latter case, we cross-validate the parameter $\rho \in \{0, 0.1, 0.2, ..., 1\}$ for each setting, but keeping it fixed across the classes.

In Tables 1, 2, 3 and 5, the bold numbers indicate the top performing representations in each setting that are statistically equivalent, which we measure by using the bootstrapping method proposed in Everingham *et al*. [14], at 95% confidence interval. In Tables 4 and 6, we are unable to run the test on other state-of-the-art approaches, as the statistical significance test requires original classification scores on the test images.

5.2 Evaluating BoW and Pólya models

In Table 1 we compare the results obtained using standard BoW histograms, two types of power normalized histograms, and the Pólya model. In all three cases, we generate the visual word counts from soft assignments of patches to the MoG components. Overall, we see that the spatial information of SPM is useful, and that larger vocabularies increase performance. We observe that both power

SPM	Method	64	128	256	512	1024
No	BoW	21.0	28.6	37.1	40.5	43.7
No	SqrtBoW	20.8	28.4	**37.6**	**41.4**	**46.0**
No	PnBoW	20.9	**30.4**	37.4	**41.5**	**46.3**
No	LatBoW	**21.7**	**30.0**	**38.4**	**41.0**	44.9
Yes	BoW	37.1	39.8	42.8	46.3	48.9
Yes	SqrtBoW	37.9	**41.3**	**44.6**	47.8	**51.6**
Yes	PnBoW	37.7	**41.4**	**44.6**	47.4	51.3
Yes	LatBoW	**39.5**	**41.8**	**45.4**	**49.2**	**52.3**

TABLE 1
Comparison of representations with and without SPM: BoW, two types of power normalized BoW, and Pólya.

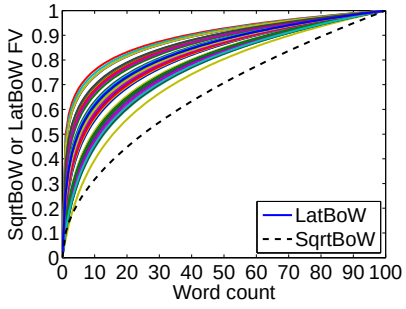


Fig. 8. Comparison of the discounting functions learned by the latent BoW model for 64 visual words (solid), and the square-root transformation (dashed). Transformed counts are rescaled to the range $[0, 1]$.

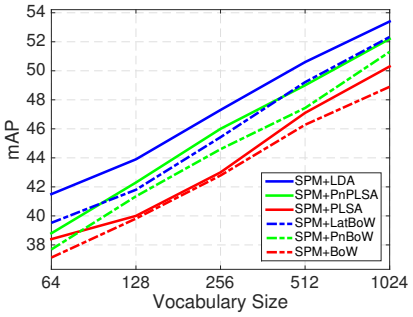


Fig. 9. Topic models ($T = 2$, solid) compared with BoW models (dashed): BoW/PLSA (red), power-normalized BoW/PLSA (green), and Pólya/LDA (blue). SPM grids are used in all experiments.

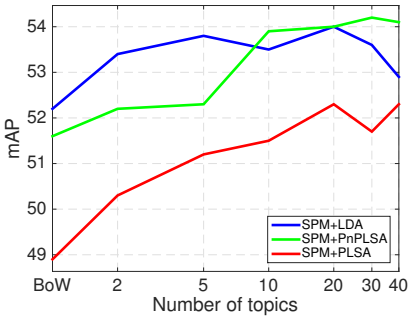


Fig. 10. Performance when varying the number of topics: PLSA (red), power-normalized PLSA (green), and LDA (blue). BoW/Pólya model performance included as the left-most data point on each curve. All experiments use SPM, and $K = 1024$ visual words.

normalization and the Pólya model both consistently improve the BoW representation, across all dictionary sizes, and with or without SPM. Furthermore, the Pólya model generally leads to larger improvements than power normalization. These results are in line with the observation of Section 4.1 that the non-iid Pólya model generates similar transformations on BoW histograms as power normalization does, and show that normalization by the digamma function is at least as effective as power normalization.

Figure 8 illustrates the discounting functions learned by the Pólya model for a dictionary of 64 visual words, without a spatial pyramid. Each solid curve in the figure corresponds to one of the visual words, and shows the corresponding digamma function $\psi(\alpha_k + n_k)$ as a function of the visual word count n_k. Compared to the square-root transformation, which is shown by the dashed curve, we observe that the Pólya model generally leads to similar but somewhat stronger discounting effect.

5.3 Evaluating topic model representations

We compare different topic model representations of Section 4.2: Fisher vectors computed on the PLSA model, its power normalized version, and using the corresponding LDA latent variable model. We compare to the corresponding BoW representations, and include SPM in all experiments. For the sake of brevity, we report only cross-validation based power normalization, as square-rooting gives similar results. In order to train LDA models, we first train a PLSA model, and then fit Dirichlet priors on the topic-word and document-topic distributions as inferred by PLSA.

In Figure 9, we consider topic models using $T = 2$ topics for various dictionary sizes, and in Figure 10 we use dictionaries of $K = 1024$ visual words, and consider performance as a function of the number of topics.

We observe that (i) topic models consistently improve performance over BoW models, and (ii) the plain PLSA representations are consistently outperformed by the power normalized version, and the LDA model. The LDA model requires less topics than (power-normalized) PLSA to obtain similar performance levels. This is in line with our findings with the BoW model of the previous section.

5.4 Evaluating latent MoG model

We now turn to the evaluation of the MoG-based image representations. In order to speed-up the learning of the hyper-parameters, we fix the patch-to-word soft-assignments as obtained from the MoG dictionary, and pre-compute the sufficient statistics of Eq. (35) once. We then iteratively update the model hyper-parameters, and the parameters of the posteriors on the per-image latent MoGs, as detailed in Section 4.3.

We initialize the Dirichlet distribution on the mixing weights by matching the moments of the distribution of normalized visual word frequencies s_k^0, which gives an approximate maximum likelihood estimation [34]. Similarly, we initialize the hyper-parameters a_k and b_k of the Gamma prior on the precision of visual word k, by matching the mean and variance of empirical precision values computed from the sufficient statistics for each visual word, while weighting the contribution of each image by the count of visual word k in that image. In this step, the empirical precision values of visual words with few associated descriptors can become too large and may lead to poor initialization. To deal with this issue, we truncate per-image empirical precision values with respect to the corresponding global empirical precision values scaled by a

SPM	Method	32	64	128	256	512	1024
No	MoG	49.1	51.4	53.1	54.3	55.0	55.9
No	SqrtMoG	51.8	54.7	56.2	58.2	58.9	60.2
No	PnMoG	**52.6**	55.0	**56.9**	**59.0**	**60.3**	**61.1**
No	LatMoG	**52.9**	**55.9**	**56.6**	**58.6**	59.5	60.2
Yes	MoG	53.1	55.4	56.2	57.1	57.4	57.6
Yes	SqrtMoG	56.0	57.9	58.9	60.3	60.5	60.8
Yes	PnMoG	56.6	**58.4**	**59.5**	**61.1**	**61.3**	**61.8**
Yes	LatMoG	**57.3**	**58.9**	**59.4**	60.4	60.7	60.7

TABLE 2
Comparison of MoG-based FV representations: plain MoG, two types of power normalized MoG, and latent MoG.

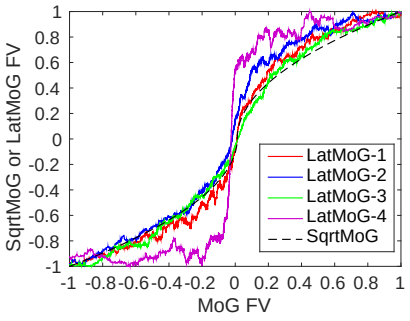


Fig. 11. Empirical comparison of components related to the Gaussian means of the power normalized MoG FVs (SqrtMoG) and latent MoG FVs (LatMoG) *vs.* the non-power-normalized FV (horizontal axis). All FV values are scaled to the range [-1,1].

constant factor, which is cross-validated among a predefined set of values. Finally, we initialize the hyper-parameters m_k and β_k by matching the mean and variance of the per-image empirical mean values computed from the sufficient statistics, again weighting each image by the count of visual word k in that image.[3]

In Table 2, we compare representations based on Fisher vectors computed over MoG models, their two power normalized versions, and the latent MoG model of Section 4.3. We can observe that the MoG representations lead to better performance than the BoW and topic model representations while using smaller vocabularies. Furthermore, the discounting effect of power normalization and our latent variable model has a more pronounced effect here than it has for BoW models, improving mAP scores by around 4 points. Also for the MoG models, our latent variable approach leads to improvements that are comparable to those obtained by power normalization. So again, the benefits of power normalization may be explained by using non-iid latent variable models that generate similar representations.

Similar to Figure 8, we present an empirical comparison of the MoG FV and LatMoG FV based on a vocabulary of size $K = 64$ components in Figure 11. In this case we consider gradients w.r.t. the Gaussian mean parameters. The transformation given by power normalization is given for reference in dashed black. Each LatMoG curve is obtained by sampling a dimension-cluster pair (d, k), and plotting the LatMoG FV with respect to $m_{k,d}$ as a function of the MoG FV with respect to $\mu_{k,d}$ over different images. The LatMoG curves are smoothed via a median filter for visualization purposes. We observe that the LatMoG model naturally generates FVs with discounting effects, as demonstrated by the curves similar to square-root transformation. Note that the gradient in Eq. (47) for the LatMoG model is a joint function of the s_k^0, s_k^1 and s_k^2 statistics, which makes that plotting LatMoG FVs against MoG FVs results in non-smooth curves.

5.5 Relationship between model likelihood and categorization performance

We have seen that the Fisher vectors of our non-iid image models provide significantly better image classification performance compared to the Fisher vectors of the corresponding iid models, unless power normalization is used to implement a discounting transformation on the image descriptors. In a broad sense, our experimental results suggest that Fisher kernels combined with more powerful generative models can possibly lead to better image categorization performance.

3. Source code for LatMoG is available at http://lear.inrialpes.fr/software.

In order to investigate the relationship between the image models and the categorization performance using the corresponding Fisher vectors, we propose to empirically analyze the MoG models and the corresponding image descriptors at a number of PCA projection dimensions (D) and vocabulary sizes (K). Here, we use the log-likelihood of each model on a validation set as a measure of the generative power of the models and evaluate the image categorization performance of the corresponding Fisher vectors in terms of mAP scores on the PASCAL VOC 2007 dataset.

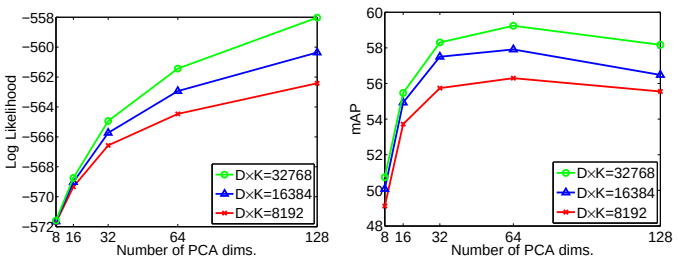
One important detail is that it may not be meaningful to compare the image categorization performance across image descriptors of different dimensionality: Our previous experimental results have shown that the mAP scores typically increase as the MoG Fisher vector descriptors become higher dimensional. Therefore, we want to compare the categorization performance across the image descriptors of fixed dimensionality, *i.e.* across the (D, K) pairs such that the product $D \times K$ is constant. On the other hand, the log-likelihood of MoG models are comparable only if they operate in the same PCA projection space. In order to overcome this difficulty, we convert each pair of PCA and MoG models into a joint generative model, which allows us to obtain comparable log-likelihood values across different PCA subspaces.

We propose to obtain the joint generative models by first defining a shared descriptor space as follows: Let $\phi(\mathbf{x}) = U^{\mathrm{T}}(\mathbf{x} - \mu_0)$ be the full-dimensional PCA transformation function for the local descriptors, where μ_0 is the empirical mean of the D_0-dimensional local descriptors and U is the $D_0 \times D_0$ dimensional matrix of PCA basis column vectors. We note that $\phi(\mathbf{x})$ does not apply dimension reduction, and the projection of a local descriptor $\mathbf{x}$ onto the D dimensional PCA subspace is given by $\mathbf{I}_{D\times D_0}\phi(\mathbf{x})$, *i.e.* the first D coordinates of $\phi(\mathbf{x})$. Therefore, the density function of a given MoG model in the D-dimensional PCA subspace is given by

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{I}_{D\times D_0}\phi(\mathbf{x}); \mu_k, \Sigma_k). \tag{49}$$

where π_k is the mixing weight, μ_k is the D-dimensional mean vector and σ_k is the variances vector of the k-th component. Then, we can map the PCA dimension reduction model and the MoG model into a new MoG model in the space of $\phi(\mathbf{x})$ descriptors as follows:

$$p_0(\mathbf{x}) = \sum_{k} \pi_k \mathcal{N}(\phi(\mathbf{x}); \mu'_k, \sigma'_k) \tag{50}$$



(a) (b)

Fig. 12. Evaluation of the model log-likelihood and the classification performance in terms of mAP scores as a function of the number of PCA dimensions (D) and the vocabulary size (K). The x-axis of each plot shows the number of PCA dimensions. Each curve represents a set of (D, K) values such that $D \times K$ stays constant.

where each mean vector is defined as

$$\mu'_k = \mathbf{I}_{D_0 \times D} \mu_k, \tag{51}$$

and each variances vector σ'_k is obtained by concatenating the corresponding D-dimensional σ_k vector with the empirical global variances of the remaining $D_0 - D$ dimensions.

In our experiments, we have randomly sampled 300,000 SIFT descriptors to measure the average model log-likelihoods. We evaluate the image categorization performance using square-rooted and ℓ_2 normalized MoG Fisher vectors, without a spatial pyramid. We have utilized (D, K) pairs obtained by varying D from 8 to 128 and K from 64 to 4096.

Figure 12a presents the model log-likelihood values and Figure 12b presents the corresponding image classification mAP scores. The x-axis of each plot shows the number of PCA dimensions. Each curve represents a set of (D, K) values where $D \times K$ is constant. From the experimental results first we can see that increasing the number of PCA dimensions (and hence reducing the number of mixing components) consistently increases the model log-likelihood. Second, the mAP scores similarly increase up to $D \leq 64$, but then degrade from $D = 64$ to $D = 128$. Therefore, even if the model log-likelihood and categorization performance are related, they are not necessarily tightly correlated. Image categorization performance can be affected by several other factors, including the details of target categorization task, and transformations applied to the Fisher vector representations, such as power and ℓ_2 normalization here. Despite these findings, we believe that further investigation of the relationship between the modeling strength of generative models and the performance of the corresponding Fisher vectors for recognition tasks can lead to advances in unsupervised representation learning.

5.6 Experiments using CNN features

We have so far utilized the SIFT local descriptors in our experiments. In this section, we evaluate the latent MoG representation based on local descriptors extracted using a convolutional neural network (CNN) model [28]. For this purpose, we consider two feature extraction schemes. First, we utilize the grid based region sampling approach based on the work by Gong *et al*. [16] and Liu *et al*. [31], and extract local descriptors by feeding cropped regions to a CNN model. Second, inspired from the R-CNN object detector [15], we propose to extract local CNN features for the image regions sampled by a candidate window generation method. Unlike the R-CNN detector, however, we utilize the region descriptors to extract image descriptors using the Fisher kernel framework, instead of evaluating individual regions as detection candidates. To the best of our knowledge, we are first to utilize detection proposals for this purpose.

Regions	CNN Layer	MoG	SqrtMoG	PnMoG	LatMoG
Grid	fc6	69.4	**74.1**	**74.3**	73.3
Grid	fc7	66.6	74.6	**75.7**	**75.3**
Selective	fc6	74.2	76.8	**77.0**	75.5
Selective	fc7	74.5	77.8	**78.0**	77.1

TABLE 3
Comparison of mAP scores on PASCAL VOC'07 dataset: plain MoG, two types of power normalized MoG and latent MoG.

Method	mAP
CNN baseline [44]	73.9
Razavian *et al*. [44]	77.2
Bilen *et al*. [1]	**80.9**
Liu *et al*. [31]	76.9
Ours (PnMoG, sel. search, fc7)	78.0
Ours (LatMoG, sel. search, fc7)	77.1

TABLE 4
Comparison of the power normalized MoG and latent MoG representations against recent results on the PASCAL VOC'07 dataset.

In order to extract CNN features from regions sampled on a grid, we follow the local region sampling approach proposed by Liu *et al*. [31]: a given image is first scaled to a size of 512×512 pixels, then, regions of size 227×227 are sampled in a sliding window fashion with a stride of 8 pixels. This procedure results in around 1300 regions per image. The image patch corresponding to each region sample is cropped and feed into the CNN model of Krizhevsky *et al*. [28], which is pre-trained on the ImageNet ILSVRC2012 dataset [11] using the Caffe library [23]. Finally, the outputs of the CNN model are used as the local descriptors.

In our second approach, we utilize the detection proposal regions generated using the selective search method of Uijlings *et al*. [50]. This method computes multiple hierarchical segmentation trees for a given image, and takes the segment bounding boxes as the detection proposals. This procedure results in around $1,500$ regions per image. Following the R-CNN object detector, we crop and re-size the window proposals to regions of size 224×224, as required by the CNN model.

As region descriptors we consider the layer six and seven activations of the CNN model. In order to speed up the Fisher vector computations, we project the original $4,096$-dimensional feature vectors to 128 dimensions using PCA. In our preliminary experiments, we have verified that higher dimensional PCA projections does not improve the image categorization performance. Following the iid MoG based experiments in [16] and [31], we use models with $K = 100$ Gaussian components, ℓ_2 normalize the resulting image representations, and do not use SPM grids.

In Table 3, we compare the MoG Fisher vector, its power normalized versions, and the latent MoG Fisher vector representations. First, we observe that using selective search regions for descriptor pooling leads to consistently better results than using the grid based regions. Given that both approaches use a comparable number of regions, the improvement using selective

Regions	CNN Layer	MoG	SqrtMoG	PnMoG	LatMoG
Grid	fc6	60.1	66.0	**67.3**	62.2
Grid	fc7	57.0	**64.8**	**65.0**	61.5
Selective	fc6	66.6	**69.4**	**69.7**	68.2
Selective	fc7	65.2	**69.0**	**69.1**	**69.1**

TABLE 5
Comparison of classification accuracy on MIT Indoor: plain MoG, two types of power normalized MoG and latent MoG.

Method	Accuracy
Juneja *et al*. [25]	63.2
Doersch *et al*. [12]	64.0
CNN baseline [44]	58.4
Razavian *et al*. [44]	69.0
Liu *et al*. [31]	68.2
Gong *et al*. [16]	68.9
Ours (PnMoG, sel. search, fc7)	**69.1**
Ours (LatMoG, sel. search, fc7)	**69.1**

TABLE 6
Comparison of the power normalized MoG and latent MoG representations against recent results on the MIT Indoor dataset.

search regions is probably due to using regions of multiple scales, and having a better object-to-clutter ratio. Second, we observe that also in this setting using the Latent MoG model leads to improvements that are comparable to those obtained by power normalization. Third, best results are obtained with layer seven activations using power normalization and our latent model.

In Table 4 we show that our results are comparable to the recent results based on a similar CNN models. The first row shows the CNN baseline (73.9%), as reported by Razavian *et al*. [44], which corresponds to training an SVM classifier over the full image CNN descriptors. The same paper also shows that the performance can be improved to 77.2% by applying feature transformations to image descriptors and incorporating additional training examples via transforming images. Bilen *et al*. [1] (80.9%) explicitly localizes object instances in images using an iterative weakly supervised localization method. The result shows that explicit localization of the objects can help better categorization of the images. Liu *et al*. [31] (76.9%) extract Fisher vectors of a sparse coding based model over local CNN features (see Appendix A for a detailed discussion of their model). Overall, we observe that our results using power normalized MoG FVs (78.0%) and latent MoG FVs (77.1%) are comparable to the aforementioned recent results, all of which are based on similar CNN models, and validate the effectiveness of our Latent MoG model for local feature aggregation.

We note that better results on the VOC'07 dataset have recently been reported based on significantly different CNN features and/or architectures. For example, Chatfield *et al*. [7] achieve 82.4% mAP by utilizing the *OverFeat* [47] architecture, combined with a carefully selected set of data augmentation, data normalization and CNN fine-tuning techniques. Wei *et al*. [54] achieve 85.2% by max-pooling the class predictions over candidate windows, utilizing additional training images, and using a two-stage CNN fine-tuning approach. Simonyan and Zisserman [48] report that the classification performance can be improved up to 89.7% mAP by using very deep network architectures, and combining multiple CNN models. We can expect similar improvements in the feature aggregation methods, including ours, by utilizing these better-performing CNN features.

In order to validate our results on a second dataset, we evaluate our latent MoG model on the MIT Indoor dataset. The dataset contains 6,700 images, each of which is labeled with one of the 67 indoor scene categories. Before extracting window proposals, we resize each image such that the larger dimension is 500 pixels. We use the standard split for the dataset, which provides 80 train and 20 test images per class, and evaluate the results in terms of average classification accuracy.

The results for MIT Indoor are presented Table 5. In each row, we evaluate a combination of the 6-th or 7-th CNN layer with the grid based or selective search based regions. Our results are overall consistent with those we obtain on VOC 2007: (i) using selective search regions leads to better performance, and (ii) using the Latent MoG model leads to significant improvements, comparable to those obtained by power normalization. Therefore, the results again support that the benefits of power normalization can be explained by their similarity to non-iid latent variable models that generate similar transformations.

Finally, in Table 6, we compare our results on the MIT Indoor dataset with the state-of-the-art. The first methods, Juneja *et al*. [25] (63.2%) and Doersch *et al*. [12] (64.0%), extract mid-level representations by explicitly localizing discriminative image regions. In the next two rows, we observe that the CNN baseline improves from 58.4% to 69.0% using the feature and image transformations proposed by Razavian *et al*. [44]. The sparse coding Fisher vectors proposed by Liu *et al*. [31] result in a comparable performance at 68.2%. Gong *et al*. [16] (68.9%) utilizes power normalized VLAD features over the CNN descriptors extracted from multi-scale grid-based regions, in combinations with the full image CNN features. Overall, we observe that our approach using power normalized MoG FVs (69.1%) and latent MoG FVs (69.1%) over selective search regions provide state-of-the-art performance on the MIT Indoor dataset.

6 Conclusions

In this paper we have introduced latent variable models for local image descriptors, which avoid the common but unrealistic iid assumption. The Fisher vectors of our non-iid models are functions computed from the same sufficient statistics as those used to compute Fisher vectors of the corresponding iid models. In fact, these functions are similar to transformations that have been used in earlier work in an ad-hoc manner, such as the power normalization, or signed-square-root. Our models provide an explanation of the success of such transformations, since we derive them here by removing the unrealistic iid assumption from the popular BoW and MoG models. Second, we have shown that gradients of the variational free-energy bound on the log-likelihood gives exact Fisher score vectors as long as the variational posterior distribution is exact. Third, we have shown that approximate Fisher vectors for the proposed latent MoG model can be successfully extracted using the variational Fisher vector framework. Finally, we have shown that the Fisher vectors of our non-iid MoG model over CNN region descriptors extracted on selectively sampled windows lead to image categorization performance that is comparable or superior to that obtained with state-of-the-art feature aggregation representations based on iid models.

APPENDIX
A. VARIATIONAL FISHER KERNEL EXAMPLES

In this section, we give two examples that illustrate applications of the variational FV framework, in addition to the models considered in the main text.

In our first example, we derive a fast variant of the MoG FV representation using the variational Fisher kernel formulation. Recall that the final MoG FV image representation is obtained by aggregating $K(1+2D)$-dimensional per-patch FVs. Therefore, the cost of feature extraction grows linearly with respect to K, D and N. One way to speed up this process, without sacrificing the descriptor dimensionality, is to hard-assign each local descriptor to visual word with the highest posterior probability. Using hard-assignment, each local descriptor produces a $(1+2D)$ dimensional descriptor, therefore, the aggregation speeds-up by a factor of K. As noted in [45], the MoG FV descriptor in this case can be also interpreted as a generalization of the VLAD descriptor [22].

Although the hard-assignment method can provide significantly speeds up in the descriptor aggregation process, it may also cause significant information loss [51]. This problem can be addressed by utilizing *clipped* posterior weights within the variational FV framework. More specifically, we can define the family of approximate posteriors $\mathcal{Q}$ as those distributions with at most K' non-zero values. The best approximation to a given $p(k|x)$ is then obtained by re-normalizing the largest K' values of $p(k|x)$ and setting the other values to zero. In this case, each patch yields a descriptor with at most $K'(1+2D)$ non-zero values, which translates into a aggregation speed up of factor $\frac{K}{K'}$. The number of non-zeros K' can be set to strike a balance between the information loss and the aggregation cost. This shows that clipping the posteriors to speed-up the computation of FVs, as *e.g.* done in [9], can be justified in the variational framework. The MoG model can also be learned in a coherent manner, by optimizing the obtained variational bound instead of the log-likelihood. This forces the MoG components to be more separated, so that the true posteriors will concentrate on few components.

As a second example, we show that the derivation of the sparse coding FVs of Liu *et al*. [31], which we have experimentally compared against in Section 5.6, can be significantly simplified using the variational formulation. In their approach, a D-dimensional local descriptor $\mathbf{x}$ is modeled by a mixture of basis vectors:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{u};\mathrm{B})p(\mathbf{u})d\mathbf{u} \tag{52}$$

where $\mathbf{u}$ is the latent vector of mixing weights of length K, and B is the dictionary matrix with each of the K columns corresponding to a D-dimensional basis vector. The distribution $p(\mathbf{x}|\mathbf{u};\mathrm{B})$ is a Gaussian with mean $\mathrm{B}\mathbf{u}$, and covariance matrix equal to a multiple of the identity matrix, and $p(\mathbf{u})$ is the Laplacian prior on the mixture weights. Liu *et al*. [31] propose to approximate $p(\mathbf{x})$ by the point estimate for $\mathbf{u}$ that maximizes the likelihood:

$$p(\mathbf{x}) \approx p(\mathbf{x}|\mathbf{u}^\star;\mathrm{B})p(\mathbf{u}^\star) \tag{53}$$

where

$$\mathbf{u}^\star = \arg\max_{\mathbf{u}} p(\mathbf{x}|\mathbf{u};\mathrm{B})p(\mathbf{u}). \tag{54}$$

In order to compute FVs for this model, we need to compute the gradients of Eq. (53) with respect to the dictionary matrix B. However, as noted in [31], this is leads to a relatively complicated calculation since $\mathbf{u}^\star$ is dependent on B. Using a series of techniques, it is shown in [31] that the gradient is given by:

$$\frac{\partial \log p(\mathbf{x})}{\partial \mathrm{B}} = (\mathbf{x} - \mathrm{B}\mathbf{u}^\star)\mathbf{u}^\star \tag{55}$$

Instead, we can use the variational Fisher kernel formulation to achieve the same result in a simpler way. For this purpose, we define the class of approximate posteriors $\mathcal{Q}$ as the set of delta peaks that put all mass at a single value $\mathbf{u}$. It is then easy to see that the optimal $q \in \mathcal{Q}$ that maximizes the variational bound is $q(\mathbf{u}^\star) = 1$ and $q(\mathbf{u} \neq \mathbf{u}^\star) = 0$. Given the optimal q, the variational FV is given by:

$$\frac{\partial F}{\partial \mathrm{B}} = \frac{\partial \mathbb{E}_q[\ln p(\mathbf{x},\mathbf{u})]}{\partial \mathrm{B}} \tag{56}$$

$$= \frac{\partial \ln p(\mathbf{x},\mathbf{u}^\star)}{\partial \mathrm{B}} \tag{57}$$

Compared to Eq. (54), this is a much simpler derivative operation since the gradient is now decoupled from the $\mathbf{u}^\star$ estimation step. It can be easily shown that the resulting gradient is equivalent to Eq. (55). This shows that the variational FV formulation can be preferable over the original FV formulation.

Acknowledgements. This work was supported by the European integrated project AXES and the ERC advanced grant ALLEGRO.

REFERENCES

[1] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, 2014.

[2] C. Bishop. *Pattern recognition and machine learning*. Spinger-Verlag, 2006.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] G. Chandalia and M. Beal. Using Fisher kernels from topic models for dimensionality reduction. In *NIPS Workshop on Novel Applications of Dimensionality Reduction*, 2006.

[5] J. C. Chappelier and E. Eckard. PLSI: The true Fisher kernel and beyond. In *Machine Learning and Knowledge Discovery in Databases*, pages 195–210. Springer, 2009.

[6] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Image categorization using Fisher kernels of non-iid image models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[9] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with Fisher vectors. In *International Conference on Computer Vision*, 2013.

[10] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Int. Workshop on Stat. Learning in Computer Vision*, 2004.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[12] C. Doersch, A. Gupta, and A. A. Efros. Mid-level Visual Element Discovery as Discriminative Mode Seeking. In *Advances in Neural Information Processing Systems*, pages 494–502, 2013.

[13] C. Elkan. Deriving TF-IDF as a Fisher kernel. In *String Processing and Information Retrieval*, 2005.

[14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge. *International Journal on Computer Vision*, 111(1):98–136, Jan. 2015.

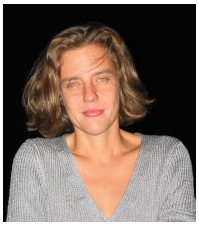
[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[16] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In *European Conference on Computer Vision*, 2014.

[17] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.
[18] T. Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems*, pages 914–920, 1999.
[19] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.
[20] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, 1999.
[21] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[22] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.
[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
[24] M. Jordan, Z. Ghahramani, T. Jaakola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.
[25] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[26] T. Kobayashi. Dirichlet-based histogram feature transform for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[27] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with Fisher vectors for image categorization. In *International Conference on Computer Vision*, 2011.
[28] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.
[29] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *Image and Vision Computing*, 27(5):523–534, 2009.
[30] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
[31] L. Liu, C. Shen, L. Wang, A. van den Hengel, and C. Wang. Encoding high dimensional local features by sparse coding based Fisher vectors. In *Advances in Neural Information Processing Systems*, 2014.
[32] D. J. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.
[33] R. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *International Conference on Machine Learning*, 2005.
[34] T. Minka. Estimating a Dirichlet distribution. http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf, 2012.
[35] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic. Free energy score space. In *Advances in Neural Information Processing Systems*, 2009.
[36] A. Perina and N. Jojic. Capturing spatial interdependence in image features: the counting grid, an epitomic representation for bags of features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.
[37] A. Perina, M. Kesa, and M. Bicego. Expression microarray data classification using counting grids and Fisher kernel. In *IAPR International Conference on Pattern Recognition*, 2014.
[38] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
[39] F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[40] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, 2010.
[41] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[42] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van-Gool. Modeling scenes with local descriptors and latent aspects. In *International Conference on Computer Vision*, 2005.
[43] A. Rana, J. Zepeda, and P. Perez. Feature learning for the image retrieval task. In *Asian Conference on Computer Vision Workshop on Feature and Similarity Learning for Computer Vision*, 2014.
[44] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *arXiv:1403.6382*, Mar. 2014.
[45] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *International Journal on Computer Vision*, 105(3):222–245, 2013.
[46] J. Sánchez and J. Redolfi. Exponential family Fisher vector for image classification. *Pattern Recognition Letters*, 59:26 – 32, 2015.
[47] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*, April 2014.
[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository*, 1409.1556, 2014.
[49] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, 2003.
[50] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal on Computer Vision*, 104(2):154–171, 2013.
[51] J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.
[52] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
[53] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2008.
[54] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. CNN: single-label to multi-label. *Computing Research Repository*, 1406.5726, 2014.
[55] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *International Conference on Computer Vision*, 2005.
[56] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal on Computer Vision*, 73(2):213–238, 2007.

Ramazan Gokberk Cinbis graduated from Bilkent University, Turkey, in 2008, and received an M.A. degree in computer science from Boston University, USA, in 2010. He was a doctoral student in the LEAR team, at INRIA Grenoble, France, from 2010 until 2014, and received a PhD degree in computer science from Université de Grenoble, France, in 2014. His research interests include computer vision and machine learning.

Jakob Verbeek received a PhD degree in computer science in 2004 from the University of Amsterdam, The Netherlands. After being a post-doctoral researcher at the University of Amsterdam and at INRIA Rhône-Alpes, he has been a full-time researcher at INRIA, Grenoble, France, since 2007. His research interests include machine learning and computer vision, with special interest in applications of statistical models in computer vision.



Cordelia Schmid holds a M.S. degree in computer science from the University of Karlsruhe and a doctorate from the Institut National Polytechnique de Grenoble. She is a research director at INRIA Grenoble where she directs the LEAR team. In 2006 and 2014, she was awarded the Longuet-Higgins prize for fundamental contributions in computer vision that have withstood the test of time. In 2012, she obtained an ERC advanced grant. She is a fellow of IEEE.